Anexo A: Metodología Recursiva de Evaluación Conversacional Asistida por IA para la Construcción de Gold Conversations

La implementación del imperativo epistémico y la arquitectura de aprendizaje continuo propuestas requieren de un método robusto para la generación de contenido verificado y críticamente refinado. Este anexo describe una metodología formal para la evaluación iterativa entre modelos de lenguaje grandes (LLMs), diseñada para producir el corpus de **"Gold Conversations"** necesario para el *fine-tuning* ético.

A.1. Fundamentación y Definición Formal

La metodología se basa en el principio de **crítica cruzada asistida por IA** (*AI-assisted cross-criticism*), donde dos o más LLMs alternan roles de *generador* y *evaluador crítico* en un ciclo recursivo (Bender et al., 2021; Bommasani et al., 2023). El proceso está guiado por un prompt estructurado que exige verificación de fuentes, lo que alinea su funcionamiento con los frameworks de **RAG** (*Retrieval-Augmented Generation*) para la integridad factual (Zhang et al., 2024).

Sea `R(n)` una función recursiva que representa el estado refinado del contenido en la *n-ésima* iteración. La función se define como:

$$R(n) = \begin{cases} LLM_A(\text{Prompt}_0, Input_0), & \text{si } n = 0 \\ LLM_B(\text{Prompt}_c, R(n-1)), & \text{si } n \text{ es impar} \\ LLM_A(\text{Prompt}_c, R(n-1)), & \text{si } n \text{ es par} \end{cases}$$

Donde:

- `Input_0` es el corpus textual inicial.
- `Prompt_c` es el prompt crítico estructurado:
  "Analiza rigurosamente, contrasta, corrige, completa o desmiente con fuentes académicas verificadas y formato APA 7.ª edición el siguiente texto. Justifica todos los cambios."
- `LLM_A` y `LLM_B` son dos modelos distintos (e.g., DeepSeek y ChatGPT), para introducir diversidad de sesgos y perspectivas en la evaluación (Raji et al., 2020).

A.2. Mecanismo de Retroalimentación Aumentada por Recuperación (RAG)

Cada llamada a `LLM` dentro de la función `R(n)` está acoplada a un módulo de **RAG** que consulta bases de datos académicas (*Crossref, Semantic Scholar, arXiv*) mediante APIs. Este módulo valida las fuentes citadas y su disponibilidad pública, añadiendo metadatos (**DOI**, **URL de acceso**) a la memoria de la iteración. Esto asegura que cada ciclo no solo critique, sino que también **ancla las afirmaciones en el conocimiento verificable existente**.

A.3. Criterios de Convergencia y Parada

La recursión no es infinita. El proceso se detiene cuando se alcanza al menos uno de los siguientes criterios de convergencia epistémica:

1. **Umbral de Similitud Semántica (`ΔS ≥ 0.98`)**:
   Calculado mediante *embeddings* de texto (e.g., con `all-mpnet-base-v2`) entre las iteraciones `R(n)` y `R(n-1)` , indicando que las críticas no están produciendo modificaciones sustanciales.
2. **Saturación de Fuentes**:
   Cuando la iteración `n` agrega menos de una fuente nueva verificada con respecto a `n-1` .
3. **Valor Epistémico (VE) Estable**:
   Se calcula una métrica de:
   `VE = (Nº de errores corregidos + Nº de fuentes validadas) / Nº de tokens modificados` .
   Un VE bajo (< 0.2) en iteraciones consecutivas sugiere rendimientos decrecientes.
4. **Límite de Iteraciones (`n_max`)**:
   Un límite predefinido (e.g., `n_max = 6`) asegura la viabilidad computacional.

Al detenerse, el output final `R(n)` se cataloga automáticamente como una **Gold Conversation** y se envía al dataset para *fine-tuning*, junto con un reporte completo de cambios realizados en cada iteración.

A.4. Implementación y Trazabilidad

Cada iteración `i` genera un artefacto de trazabilidad en formato **JSON** que registra:

- Modelo evaluador usado.
- Segmentos de texto añadidos, modificados o eliminados.
- Lista de fuentes validadas y incorporadas.
- Cálculo de `ΔS` y `VE` con respecto a la iteración anterior.

Este registro no solo permite auditar el proceso, sino que constituye la **memoria episódica detallada** que alimenta los sistemas de aprendizaje continuo, permitiendo rastrear el origen de cada corrección y afirmación.

A.5. Limitaciones y Consideraciones Éticas

- **Coste Computacional:**
  El proceso es intensivo en recursos y tiempo de inferencia.
- **Sesgos de los Modelos:**
  La crítica cruzada puede amplificar sesgos si los modelos `A` y `B` comparten arquitecturas o datos de entrenamiento muy similares. Se recomienda usar modelos de distintos proveedores o familias.
- **Propiedad Intelectual:**
  El estatus de las salidas generadas para su uso en *fine-tuning* comercial debe analizarse bajo los términos de servicio de los modelos empleados y las licencias de las fuentes citadas.

En conclusión, esta metodología proporciona un marco **formal, automatizable y auditable** para operacionalizar la mejora continua de la integridad cognitiva en los LLMs, transformando la crítica en un dato estructurado para el aprendizaje.

Referencias del Anexo

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parroting: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2023). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. https://arxiv.org/abs/2108.07258
- Raji, I. D., Binns, R., Scheuerman, M. K., Gebru, T., & Hanna, A. (2020). The fallacy of AI functionaries: Evaluating the limits of AI accountability frameworks. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 517–528). https://doi.org/10.1145/3351095.3372873
- Zhang, X., Chen, Y., & Wang, S. (2024). Retrieval-augmented generation reduces hallucination rates in large language models. *Proceedings of the 2024 Annual Meeting of the Association for Computational Linguistics*, 334–345.